WHAT IS CLAIMED IS:

1. A method of processing a client request for a file, comprising:

5   transmitting a first fragment of the file that is stored in a first tier of server storage to the

       client;

       retrieving a subsequent fragment of the file from a lower tier of storage while the first

       fragment is transmitting; and

10

       after transmission of the first fragment completes, transmitting the subsequent fragment

       to the client.

2. The method of claim 1, wherein transmitting the first fragment includes retrieving the first

15 fragment from a file cache of the server.

3. The method of claim 2, wherein the file cache includes a first portion in which the first

fragment is stored, and further comprising storing the subsequent fragment in a second portion of

the file cache.

20

4. The method of claim 2, wherein the file cache comprises a portion of the volatile system

memory of the server.

5. The method of claim 1, wherein the lower tier of storage comprises at least one of a server

25 disk device, a networked storage device, or a remote system memory.

6. The method of claim 1, further comprising, responsive to determining that a first fragment of

the requested file is not valid in the first tier of storage, retrieving the first fragment from a lower

tier of storage and storing the first fragment in the first tier.

30

7. The method of claim 6, further comprising determining a size for the first fragment based upon the transmission window of a connection between the server and client.

8. The method of claim 7, wherein the first fragment size is less than or equal to the maximum active transmission window of the server.

9. The method of claim 1, wherein transmitting the first fragment includes formatting the first fragment according to the transmission control protocol (TCP).

10. A server device, comprising:

     a processor;

     a system memory accessible to the processor and configured with instructions suitable for execution by the processor;

     server code means for transmitting a first fragment of the file that is stored in a first tier of server storage to the client;

     server code means for retrieving a subsequent fragment of the file from a lower tier of storage while the first fragment is transmitting; and

     server code means for transmitting the subsequent fragment to the client after transmission of the first fragment completes.

11. The server device of claim 10, wherein the code means for transmitting the first fragment includes code means for retrieving the first fragment from a file cache of the server.

12. The server device of claim 11, wherein the file cache includes a first portion in which the first fragment is stored, and further comprising code means for storing the subsequent fragment in a second portion of the file cache.

13. The server device of claim 11, wherein the file cache comprises a portion of the server system memory.

14. The server device of claim 10, wherein the lower tier of storage comprises at least one of a server disk device, a networked storage device, or a remote system memory.

15. The server device of claim 10, further comprising, code means for retrieving the first fragment from a lower tier of storage and storing the first fragment in the first tier responsive to determining that a first fragment of the requested file is not valid in the first tier of storage.

16. The server device of claim 15, further comprising code means for determining a size for the first fragment based upon the transmission window of a connection between the server and client.

17. The server device of claim 16, wherein the first fragment size is less than or equal to the maximum active transmission window of the server.

18. The server device of claim 10, wherein transmitting the first fragment includes formatting the first fragment according to the transmission control protocol (TCP).

19. A computer program product residing on a computer readable medium for enabling a server device to process client requests, comprising:

   server code means for transmitting a first fragment of the file that is stored in a first tier of server storage to the client;

server code means for retrieving a subsequent fragment of the file from a lower tier of storage while the first fragment is transmitting; and

server code means for transmitting the subsequent fragment to the client after transmission of the first fragment completes.

20. The computer program product of claim 19, wherein the code means for transmitting the first fragment includes code means for retrieving the first fragment from a file cache of the server.

21. The computer program product of claim 20, wherein the file cache includes a first portion in which the first fragment is stored, and further comprising code means for storing the subsequent fragment in a second portion of the file cache.

22. The computer program product of claim 20, wherein the file cache comprises a portion of volatile server system memory.

23. The computer program product of claim 19, wherein the lower tier of storage comprises at least one of a server disk device, a networked storage device, or a remote system memory.

24. The computer program product of claim 19, further comprising, code means for retrieving the first fragment from a lower tier of storage and storing the first fragment in the first tier responsive to determining that a first fragment of the requested file is not valid in the first tier of storage.

25. The computer program product of claim 24, further comprising code means for determining a size for the first fragment based upon the transmission window of a connection between the server and client.

26. The computer program product of claim 25, wherein the first fragment size is less than or equal to the maximum active transmission window on the server.